Literature Review: How Question-Answering Models Handle Unanswerable Queries

Samuel Schreiber,
srs17@illinois.edu

December 5, 2024

Abstract

This review covers four papers that discuss how question-answering models are trained and evaluated on unanswerable questions. It seeks to define what an unanswerable question is, how models can handle them without hallucination, and how performance can be evaluated. It covers specific types of unanswerable questions, including those with 'questionable assumptions' in an open-domain setting, plausible unanswerable questions in a closed-domain setting where the question cannot be answered based on some provided context, and adversarial attacks where the question is deliberately crafted to trick the model into providing a bogus answer. Finally, it examines information-seeking questions where the answer could be found across many contexts, and the subtasks of paragraph retrieval and answerability detection.

1 Introduction

Question-Answering models have made significant progress in recent years, with models like GPT-3 and BERT achieving human-level performance on many question-answering tasks. Performance for these models is often evaluated over good-faith, well-formulated questions with answers that are present either in an **open or closed domain setting** (see 4.2.1). In some cases, such as in a closed-domain setting where users ask questions based on specific contexts (such as an article), models that appear to perform well may simply be matching questions to the **context span** (or snippet) that is most related to the question. For example, if the context is *'... The age of the Earth is 4.543 billion years old. ...'*, to answer the question of *'How old is the Earth?'*, the model may simply match the question to the context span *'Earth is 4.543 billion years old'*. In such case, there is no reason to believe the model has any true natural language understanding of the question; it could simply be pattern-matching based on some similarity metric. Moreover, the context span may simply refer to the entities or concepts in the question without providing any meaningful answer. For this reason, it is important to reassess how models are trained and evaluated so that they can handle unanswerable questions, provide more meaningful answers beyond matching the question to a context span, and provide more reasonable performance metrics more reflective of the model's performance in the real world. But what is an unanswerable question, and how can models learn to detect them to provide a reasonable response rather than hallucinating (or making up some spurious) answers? Finally, how can we evaluate the performance of models on such questions? These are the questions we seek to clarify in this review.

2 Know what you don't know: Unanswerable questions

2.1 Motivation

This paper (Rajpurkar et al., 2018) was selected because it points out a common flaw in evaluating question-answering models and that many models only perform well because the question-answering datasets only contain answerable questions or easily identifiable unanswerable questions, resulting in a biased view of the models' performance. It raises awareness about the need to think critically about how models are evaluated and how biases that may be imposed on the test dataset may affect the models' performance. This paper is relevant to our overarching question because it identifies a new type of unanswerable question, and offers a new dataset to train and evaluate models on unanswerable questions to avoid hallucinating an answer.

2.2 Background Knowledge

2.2.1 SQuAD 1.1 Dataset

This paper discusses **SQuAD** (Stanford Question-Answering Dataset), a commonly used dataset used to train and evaluate question-answering models. It consists of many question-answer pairs, where the model is provided a context with a set of questions and is expected to provide the respective answers. The models' answers are compared to the ground truth answers provided in the dataset to determine if the answer was relevant or not (see the section 2.2.2 for more information about model evaluation). It is vital to note that the dataset only contains answerable questions.

Here is an example of a question from the SQuAD dataset:

> **Context:** The Normans were originally Viking raiders from Scandinavia who settled in the region of France known as Normandy. They became known for their military prowess and eventually played a significant role in the history of England, particularly after their victory in the Battle of Hastings in 1066.
>
> **Question 1:** Where did the Normans settle in France?
> Answer: Normandy
>
> **Question 2:** When was the Battle of Hastings?
> **Answer:** 1066

Notice the context provides a set of facts, and all the question's answers can be found in the context in contrast to section 4 where specific context is not provided. Humans may manually determine if the answer a model generates was correct based on the context, or determine if the question was correctly labeled as unanswerable.

2.2.2 F1 Score

A common way to evaluate the performance of a question-answering model is to calculate the **F1 score** which is a harmonic mean of the **precision and recall** of the models' answers. In this case, the precision is the percentage of tokens (often words or sub-words) in the models' answer that are also in the ground truth answer, and the recall is the percentage of tokens in the ground truth answer

that are also in the models' answer. Moreover, precision tells us how much of the generated answer is correct, and recall tells us about the coverage of the generated answer.

2.2.3 Development, Test, and Training Sets

The SQuAD dataset is split into three sets: **the training set, the development set, and the test set**. The training set is used to train the model, the development is used during training to tune the model's hyperparameters and check the model's performance during training, and the test set is used to evaluate the model's performance after the model's training has finished. It is essential that the final model not be trained on the development or test sets to avoid invalid performance evaluation.

2.3 Overview

The paper discusses how the SQuAD 1.1 dataset only contains answerable questions and models that perform well may not have any true understanding of the question. Models that perform well may simply match a question to the span of the context that is most related to the question rather than understanding the question and providing a meaningful answer. To address this shortcoming, the authors propose **SQuAD 2.0**, which augments the original dataset with over 50,000 unanswerable questions. The authors show that state-of-the-art question-answering models in 2018, at the time of the paper's publication, achieve an F1 score of 66.3% on SQuAD 2.0, compared to 85.8% on SQuAD 1.1, highlighting the models' struggle with unanswerable questions. Furthermore, the performance of state-of-the-art models on SQuAD 1.1 is 5.4 points worse than humans, while the performance on SQuAD 2.0 is 23.2 points worse than humans, further demonstrating the models' shortcoming's in handling unanswerable questions.

2.3.1 Data Collection

The authors collected unanswerable questions by asking crowd workers to write unanswerable questions for selected SQuAD 1.1 contexts. They were instructed to pose five questions per context that referenced entities or concepts that were present but not directly answerable by the context. The questions must have also been plausibly answerable by a human. There were some filters in place to remove questions from humans who did not produce quality questions. These questions were then added to the SQuAD 1.1 dataset to create SQuAD 2.0.
The authors used crows workers to have humans answer questions from SQuAD 2.0 development and test sets to confirm that the dataset was clean. They were exposed to mixed sets of unanswerable and answerable questions and were asked to either underline the answer in the context or mark it as unanswerable. If more than half of workers marked a question as unanswerable, it was labelled as unanswerable. 93% of the questions that were marked as unanswerable were truly unanswerable.

2.4 Contribution, Critique and Analysis

The curation of the SQuAD 2.0 dataset is a significant contribution to question-answering research. It highlights a bias in the original SQuAD 1.1 dataset that results in less meaningful evaluation of question-answering models. It lessens the likelihood that models are simply matching questions to the context span that is related to the question and forces the models to have a deeper understanding of how the questions are related to the context. While the dataset is not perfect, containing 7% noise in the unanswerable labels due to false positives, the authors did their due diligence to ensure the integrity of the dataset to keep the noise low enough for meaningful evaluation. The exposure of

unanswerable questions reduces the likelihood that a trained model will hallucinate an answer when a question is unanswerable, assuming the model performs well on the SQuAD 2.0 test set. One limitation with this dataset that we will see in section 3 is its lack of adversarial examples, ones that are deliberately crafted to fool a model. Another limitation is that models trained on SQuAD 2.0 do not generalize well to out-of-domain samples (Sulem et al., 2021). A final limitation is this dataset is closed-domain, meaning that the answer to a question can be found within the context provided if it exists. Often times, when we think of contributions to the field of machine learning, we think of new architectures, algorithms, techniques, or equations that improve the performance of a model over some dataset. However, rather than examining the inner workings of a particular model, this paper highlights flaws of a popular dataset used to train and evaluate question-answering models and carefully curates a new version of the dataset to address them.

3 The Impacts of Unanswerable Questions on the Robustness of Machine Reading Comprehension Models

3.1 Motivation

This paper (Tran et al., 2023) was selected because it evaluates how models trained on SQuAD 2.0 and 1.1 dataset (see section 2 and section 2.2.1 respectively) perform on **adversarial attacks**, questions used to trick the model into generating incorrect answers. It serves as a natural continuation of the previous paper, highlighting strengths and limitations of the SQuAD 2.0/1.1 datasets. Ensuring models are robust against adversarial attacks is vital before deploying such models for general consumer use. Moreover, robustness to such attacks provides us with more confidence that the model is not simply exploiting statistical patterns and has a deeper understanding of our natural language. It is relevant to our overarching questions because it covers a new type of unanswerable question, adversarial attacks, and elaborates on the first paper.

3.2 Background Knowledge

3.2.1 Adversarial Attack

An **adversarial attack** in the context of MRC (Machine Reading Comprehension) is a question that is deliberately crafted to trick the model into providing a bogus answer. One such example the authors mention is one that misleads the model into an incorrect conclusion by including large lexical overlap with the source context. For example:

> **CONTEXT:** Dartfort is the name of the state that the megaregion expands to in the West.
>
> **QUESTION:** What is the name of the state that the megaregion expands to in the East?
>
> **MODEL'S ANSWER:** Dartfort

3.3 Overview

The paper evaluates the performance (see 2.2.2) of three pre-trained state-of-the-art question-answering models (at the time of the paper's publication): BERT, SpanBERT, and RoBERTa, fine-tuned on the

SQuAD 2.0 and 1.1 datasets, on adversarial attacks. The authors propose an algorithm that transforms a triplet of context, question, and answer into an adversarial triplet. They set aside an original dataset and an adversarial dataset (using the algorithm to transform the original dataset) to evaluate the models' performance. For each model, they fine-tune a copy of the model on SQuAD 2.0 and another on SQuAD 1.1, and evaluate the models' performance on the original and adversarial datasets. The models perform well on the original datasets when fine-tuned with SQuAD 2.0, and noticeably worse when fine-tuned with SQuAD 1.1. Surprisingly, when evaluated on adversarial attacks, the models fine-tuned on SQuAD 2.0 appeared to perform worse than the models fine-tuned on SQuAD 1.1 across all three models. The authors investigate this further by breaking the model responses into four different categories.

1. **I**: Incorrect. Answerable questions that the model gets wrong, or predicts unanswerable.
2. **C2I**: Correct to Incorrect. Questions that the model gets right until they are adversarially modified.
3. **C2U**: Correct to Unanswerable. Questions that the model gets right until they are adversarially modified, and the model incorrectly predicts that the question is unanswerable.
4. **C2C**: Correct to Correct. Questions that the model gets right, even after they are adversarially modified.

The results are summarized in a table taken from the original paper in Figure 1. The authors investigated the C2U case specifically by examining how well the models' second-best answers performed, as defined by the models' confidence. They found that the models' second-best answers were 'fairly good', but the models failed to put them ahead of the unanswerable response. They hypothesize that models' fine-tuned on unanswerable questions are able to perceive attacks on answerable questions, but are unable to propose the right answer with the highest confidence. The authors re-evaluated the models, eliminating the unanswerable response from SQuAD 2.0 fine-tuned models to 'force' the model to answer (see Figure 2). They found that the SQuAD 2.0 fine-tuned models' performance was noticeably better than the SQuAD 1.1 fine-tuned models on the adversarial attacked samples. Finally, the authors performed similar evaluations on many samples from out-of-domain setting, and found that models' fine-tuned on SQuAD 2.0 always had superior performance to the SQuAD 1.1 fine-tuned models on adversarial attacks.

3.4 Contribution and Analysis

This paper demonstrated that models fine-tuned on unanswerable questions, specifically from SQuAD 2.0, are more robust to adversarial attacks than models fine-tuned on SQuAD 1.1. This is significant, because even though the fine-tuned models were never exposed to any adversarial attack, they were still able to generalize better to them if they had been trained on unanswerable questions in general. One possible limitation is that the adversarial examples are generated synthetically from an algorithm rather than by humans which is an important distinction to make. Also, the paper does not specify how many adversarial examples were generated or how many original samples were used to generate them.

4 Question-Answering with Questionable Assumptions

4.1 Motivation

Previously, SQuAD 2.0 was introduced which includes unanswerable questions for better question-answering models evaluate. The next paper (Najoung Kim et al., 2023) hones in on a specific type of unanswerable question, one where the question assumes one or more facts that are false or **unverifiable**. The authors provide an example of such a question:

> When did Marie Curie discover Uranium?

This is a plausible and well-framed question. However, the question presupposes that Marie Curie discovered Uranium, which is false. Therefore, the question of 'When' is unanswerable. This paper is included to offer an approach for detecting such questions and determining an appropriate response to them that does not involve hallucinating an answer.

4.2 Background Knowledge

4.2.1 Open vs Closed Domain Question-Answering

This paper focuses on the diagnosis of questions containing unverifiable assumptions in an **open-domain setting** in contrast to the SQuAD 2.0 dataset which is considered a **closed-domain setting**. SQuAD 2.0 is closed-domain because each question is paired with a specific context, and it is assumed that the answer (if one exists) can be found within that context. An open-domain setting makes no such assumption. The answer could potentially be found across any number of documents or contexts, requiring the model to retrieve and process multiple pieces of information before providing an answer. This paper is relevant to our overarching question because it introduces a new type of unanswerable question, one with questionable assumptions, and offers a new dataset to evaluate models on such questions.

4.2.2 Questionable Assumptions

In the context of this paper, **questionable assumptions** are presuppositions baked into the question that are either demonstrably false or unverifiable such in the Marie Curie example. It is important to note that these questionable assumptions are assumed to not be deliberately crafted to confuse the model, but instead stem from genuine ignorance in contrast to section 3 where the intent is more malicious.

4.2.3 Epistemic Bias

They authors acknowledge that not all presuppositions are directly implied by the question but are reasonably safe to assume. An epistemic bias is a presupposition that is not baked into the question but is likely believed by the speaker. For example, *'How many great white sharks are in captivity?'* does not imply that there are any great white sharks in captivity, but the speaker likely believes that there are or else it would not be a reasonable question to ask. The authors refer to this as the **epistemic bias of the speaker**.

4.2.4 Wh-Questions

Questions that begin with 'Wh-' such as 'When', 'Where', 'Who', 'What', 'Why', and 'How' are called **Wh-questions**. These are the questions that the authors are concerned with in this paper.

4.2.5 Adaptation Set

An **adaptation set** is a subset of data, generally much smaller than a training set, used for fine-tuning pre-trained models rather than training them from scratch.

4.3 Overview

The authors propose a method for evaluating a model's performance on questions with questionable assumptions in an open-domain setting. In this paper, questionable assumptions include **epistemically biased propositions** (4.2.3). They construct a dataset called $(QA)^2$ containing plausible questions with questionable assumptions based on frequent search-engine queries. There are a total of 602 questions, half of which contain questionable assumptions. 32 questions out of the 602 are set aside to form an adaptation set, where half of the 32 contain questionable assumptions. While this seems like a relatively small number of questions in comparison to SQuAD 1.1/2.0, the authors make it clear that the dataset is primarily intended to be used to evaluate model performance in contrast to the SQuAD 2.0 dataset which can be used for training. To assess model performance, the authors employed crowdworkers to rate the acceptability of the model's answers to the questions. Additionally, they evaluated the models on two binary classification tasks: detecting whether a question contains questionable assumptions and determining whether a specific proposition can be supported (verified) based on the available data. They found that GPT 3 and 3.5 had a human acceptability judgement score of 54%, and a score of 68% and 72% for the verification and questionable assumption detection tasks respectively.

4.4 Data Collection

The authors used Google's autocomplete to collect their questions, focusing on **Wh-questions** (4.2.4) with *wh-* prefixes that can be autocompleted into *wh-* questions. Next, they used crowdsourcing to identify questions with questionable assumptions. Finally, expert annotators were used to verify if the questionable assumptions identified by the crowdworkers were indeed questionable assumptions.

4.5 Limitations

4.5.1 Evaluation but not Training

The dataset is primarily designed to evaluate question-answering models on 'questionable assumptions', though it does set aside an adaptation set for fine-tuning models. In contrast, the SQuAD 2.0/1.1 dataset can be used for both training and evaluation. This means that the model must learn to detect unverifiable assumptions from a different dataset than it is evaluated on.

4.5.2 Open Domain Only

Another limitation is that this dataset may only effectively evaluate models that are trained over an open-domain setting such as GPT. It would not be effective at evaluating models that are trained in

a closed-domain setting, such as over a Biology textbook.

4.5.3 Small Dataset

While the dataset was diligently curated, it is still relatively small with only 602 questions, 32 of which are used for adaptation. Out of the 602 questions, half contain questionable assumptions leaving 301 questions that contain questionable assumptions. This is in contrast to the SQuAD 2.0 dataset which contains over 50,000 unanswerable questions. While this provides a good starting point for evaluating models on such questions, it may not be enough to provide a comprehensive evaluation. Finally, though the authors made it clear the dataset is primarily for evaluation, the adaptation set is very small, and it is questionable whether it is enough to effectively fine-tune a model.

4.5.4 Biases of Search Engines

While scraping Google APIs is a clever idea to quickly collect many questions occurring in the real world, it is subject to the biases of the Google search engine which is proprietary and not well-understood. For example, Google could be amplifying certain types of questions over others, while suppressing others. Furthermore, Google may be tailoring search results to the user associated with the API key using personal information such as location, age, or interests. It is impossible to know the extent to which this is happening. The authors do not address these concerns in the paper, which could be seen as an oversight. See the Hao et al. (2020) paper for a discussion on the biases of search engines.

4.6 Contributions and Analysis

This paper is a significant contribution to the field of question-answering research in an open-domain setting. It points out that many question answering models struggle to provide adequate responses to questions with questionable assumptions, with GPT 3 and 3.5 achieving a human acceptability judgement score of just 54% on the $(QA)^2$ dataset. It proposes a method for scraping questions with questionable assumptions from search engines to ensure that the questions are plausible and have likely really been asked by human users in contrast to synthetically generated questions which may be unreasonable. While the dataset is relatively small, and it is subject to the biases of the Google search engine, in practice it may be an effective way to evaluate models on such questions as long as the limitations are kept in mind.

5 Challenges in Information-Seeking QA: Unanswerable Questions and Paragraph Retrieval

5.1 Motivation

While the previous papers focused on handling unanswerable questions given a specific context, sometimes a user provides a question before they have any target article in mind. These are called **information-seeking questions**. An example of this is a search engine that allows a user to enter a query to seek an answer based on a large set of documents known to the engine that the user may not be aware of. This paper (Asai and Choi, 2021) focuses on, among other topics, two challenges of answering such information-seeking questions: **question answerability prediction** and

target paragraph retrieval. This paper is relevant to our overarching question because it highlights the importance of question answerability prediction in information-seeking QA to improve model performance and avoid hallucinating an answer.

5.2 Background Knowledge

5.2.1 Natural Questions Dataset

Natural Questions (NQ) is an open-domain dataset containing the kinds of questions users commonly ask in real-world settings, such as through a search engine like Google. It is developed by Google. The dataset provides upper-bound performance metrics by estimating the performance of a single annotator (called **single human**), and the aggregate performance of 25 annotators (**super-human**).

5.2.2 Question Types in NQ

There are a few different types of question answers that the authors consider, all of which collectively define the question's **answerability**. A **long** answer is a paragraph that contains the answer to the question. If the answer is split between multiple paragraphs, the authors would not consider a long answer to be present. A **short** answer is the set of spans in the long answer (the paragraph) that contain the answer (think about highlighting answer-containing text). Note that the short answer is a set of spans, while the long answer is a paragraph. Finally, some questions are **unanswerable**, meaning there is no long answer (and therefore no short answer) to the question. The authors of this paper refer to long answers which have no short answer as **long only** to make it clear that the answer is present, though no short answer is present. The authors refer to the true long answers as **gold long answers** or **gold paragraphs**, and refer to the answerability of a question as its **gold type**.

5.3 Overview

The first part of the paper focuses on how question answerability prediction and target paragraph retrieval affect the performance of question-answering models. There are three different answerability types the authors consider: **short**, **long only**, and **unanswerable**. They perform an analysis on RikiNet and ETC (and a few other models this review will not cover) over the NQ dataset given the **gold paragraph** containing the answer (or gold long answer), or the **gold question type**. It is important to note that RikiNet works by first predicting the answerability type and using it as a bias for answer span prediction, while ETC jointly predicts the type and answer span. The results are shown in Figure 3 taken from the original paper, with the columns P, R, and F1 representing precision, recall, and F1 score respectively (see 2.2.2). The table illustrates that having the gold paragraph and gold answerability type are almost equally important for short-answer performance on NQ. Both models exceed the single human performance for long answer identification after they are given the gold type, and fall just short of super-human performance. For short answers, both models are improved by the gold type by around 5% F1 score. However, for ETC, being given the gold type yielded marginally better performance than when given the gold paragraph for short answers.

5.4 Contribution and Analysis

This paper yielded an outcome that may be surprising to some: that detecting the answerability of an information-seeking question is roughly as important as having the gold paragraph for short-answer performance on the NQ dataset. Or put another way, knowing whether a question is answerable is roughly as important as knowing the answer. It is important to remember that this result is specific to two models over the NQ dataset, and more research would be needed to make any larger generalizations. That said, the implications of this result are significant, and highlight the importance of question answerability prediction in information-seeking QA.

6 Conclusion

We have read and discussed four papers all of which help answer our overarching questions: what is an unanswerable question, how can models be trained and be evaluated on them, and how can they provide an appropriate response rather than hallucinating an answer? The first paper introduced **SQuAD 2.0**, a closed-domain dataset that contains many unanswerable questions that models can be trained and evaluated on. The authors introduce a specific kind of unanswerable question, one that cannot be found in a specific context. They discuss how models can be trained to detect unanswerable questions and provide a response indicating that the question is unanswerable. We discuss its limitations, including its lack of adversarial examples and closed-domain scope. The second paper evaluates how models fine-tuned on SQuAD 2.0 are more robust to a specific type of unanswerable question, **adversarial attacks**, even when not explicitly trained on them. Models are able to successfully detect many adversarial examples to avoid hallucinating answers. The third paper introduces a new dataset, $(QA)^2$, aimed at evaluating models on another type of unanswerable question: those with **unverifiable assumptions**, and those with **incorrect epistemic bias**. We discuss some limitations including the size of the dataset and how the data was collected. Finally, we discuss a paper that introduces another type of unanswerable question: **information-seeking questions**. The authors show that over the NQ dataset detecting the answerability of a question is roughly as important as knowing where the answer is. This has significant implications for how models should be trained and evaluated on information-seeking questions. In conclusion, these papers define different types of unanswerable questions and provide meaningful methods for training and evaluating models on them to avoid hallucinating answers.

7 Figures

		I	C2I	C2U	C2C
BERT	v1	10.9	28.7	-	60.4
	v2	21.3	10.9	14.7	53.2
RoBERTa	v1	8.0	24.5	-	67.7
	v2	14.5	8.0	20.5	57.1
SpanBERT	v1	8.0	26.7	-	65.4
	v2	13.8	8.3	20.1	57.8

Table 3: The percentage of answerable questions by types of answers produced by v1 and v2 models before and after adversarial attacks.

Figure 1: Taken from the original paper (Tran et al., 2023)

		Answerable		
		Original	Attacked	$\Delta \downarrow$
BERT	v1	88.4	63.8	24.6
	v2	88.5	69.6	**18.9**
RoBERTa	v1	91.5	70.5	21.0
	v2	91.4	75.1	**16.4**
SpanBERT	v1	91.5	68.6	22.9
	v2	91.3	75.8	**15.5**

Table 5: The performance of v1 and v2 models (when being forced to output non-empty answer on answerable questions) before and after adversarial attacks.

Figure 2: Taken from the original paper (Tran et al., 2023)

	Long answer			Short answer		
	P	R	F1	P	R	F1
RikiNet	74.3	76.3	75.2	61.4	57.3	59.3
w/Gold T	85.2	85.2	85.2	64.6	64.6	64.6
ETC	79.7	72.2	75.8	67.5	49.9	57.4
w/Gold T	84.6	84.6	84.6	62.5	62.5	62.5
w/Gold P	-	-	-	67.9	57.7	62.4
w/Gold T&P	-	-	-	68.9	67.6	68.3
Human						
- Single	80.4	67.6	73.4	63.4	52.6	57.5
- Super	90.0	84.6	87.2	79.1	72.6	75.7

Table 2: Oracle analysis on the dev set for NQ. "Gold T" denotes Gold Type, and "Gold P" denotes "Gold Paragraph".

Figure 3: Taken from the original paper (Asai and Choi, 2021)

References

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking qa: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 1492–1504, Online. Association for Computational Linguistics. Accessed: 2024-12-02.

Hao, Ke, et al. 2020. Algorithmic amplification of biases on google search. Accessed: 2024-12-02.

Phu Mon Htut Najoung Kim, Samuel R. Bowman, and Jackson Petty. 2023. Question-answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 8466–8487, Toronto, Canada. Association for Computational Linguistics. Accessed: 2024-12-02.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics. Accessed: 2024-12-02.

Elior Sulem, Jamaal Hay, and Roth Dan. 2021. Do we know what we don't know? studying unanswerable questions beyond squad 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, page 4543–4548, Punta Cana, Dominican Republic. Association for Computational Linguistics. Accessed: 2024-12-02.

Son Quoc Tran, Phong Nguyen-Thuan Do, Uyen Le, and Matt Kretchmar. 2023. The impacts of unanswerable questions on the robustness of machine reading comprehension models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 1543–1557, Dubrovnik, Croatia. Association for Computational Linguistics. Accessed: 2024-12-02.